

DOCUMENT RESUME

ED 453 231

TM 032 582

AUTHOR Crislip, Marian A.; Chin-Chance, Selvin
TITLE Using Traditional Psychometric Methodologies and the Rasch Model in Designing a Test.
PUB DATE 2001-04-00
NOTE 18p.; Paper presented at the Annual Meeting of the American Educational Research Association (Seattle, WA, April 10-14, 2001).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Difficulty Level; *Item Response Theory; Psychometrics; *Reliability; State Programs; *Test Construction; Testing Programs; *Validity
IDENTIFIERS Hawaii; *Rasch Model

ABSTRACT

This paper discusses the use of two theories of item analysis and test construction, their strengths and weaknesses, and applications to the design of the Hawaii State Test of Essential Competencies (HSTEC). Traditional analyses of the data collected from the HSTEC field test were viewed from the perspectives of item difficulty levels and item discrimination indices. Both classical test theory and item response theory analyses were applied to the HSTEC items. It appears that using both approaches during the test development process will make the resulting instrument more likely to provide data that are reliable for making inferences that are valid. The value that each perspective contributed toward test development also supports the idea that no one method should be used in developing tools to measure latent traits. Whether or not the item tests what is valued should be the primary emphasis in examining test items. (SLD)

Using Traditional Psychometric Methodologies and the Rasch Model in Designing a Test

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL HAS
BEEN GRANTED BY

M. Crislip

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☒ This document has been reproduced as
received from the person or organization
originating it.

☐ Minor changes have been made to
improve reproduction quality.

• Points of view or opinions stated in this
document do not necessarily represent
official OERI position or policy.

By

Marian A. Crislip, Ph.D.
Hawaii Department of Education

Selvin Chin-Chance, Ph.D.
Hawaii Department of Education

April 2001

Paper presented at the Annual Meeting of the American Educational Research
Association, Seattle, WA

BEST COPY AVAILABLE

Using Traditional Psychometric Methodologies and the Rasch Model in Designing a Test

Background

The Hawaii State Tests of Essential Competencies (HSTEC) was a minimal competency test administered to students in Hawaii public schools for certification for graduation. It was administered five times over three years and had accommodations available for students having special needs. The test consisted of two subtests: basic skills and other life skills that are subdivided further into sixteen essential competencies. These subtests, minimum pass scores and maximum scores are outlined in Figure 1:

Insert Figure 1 here.

To alleviate bias associated with test-retest measures the Test Development Section of the Hawaii State Department of Education updates the items on the HSTEC. For the field test, items were recycled from previous editions.

The 60 item field test was administered to 836 grade 10 students from seven high schools across the state. The field test was administered to the students during their social studies class with the knowledge that it was to test items for future use in the state testing program. Although the sample's ethnic distributions did not mirror that of the state, it did represent it somewhat closely.

Each of the six subtests (reading, math, reasoned solutions, fact/opinion, health, and global diversity) had 10 items. School aggregated socioeconomic status (SES), based on federal lunch subsidies, varied as well as grade level organization (e.g. K-12, 7-12, 9-12); participation in the field test was voluntary. Nunnally (1967) advised that sample size rule of thumb for such enterprises should have 5 to 10 times as many subjects as items.

The purpose of this paper is to discuss the use of two theories of item analyses and test construction, their strengths and weaknesses and applications to the design of the Hawaii State Test of Essential Competencies (HSTEC).

Methodology

Traditional Psychometric Methodologies

Item analysis is a broad term used to define the computation and examination of any statistical property of examinees' responses to an individual test item. Item parameters typically fall into three categories:

- indices that describe the distribution of responses to a single item (i.e. the mean and variance of the item responses),

- indices that describe the degree of relationship between response to the item and some criterion of interest,
- indices that are a function of both item variance and relationship to a criterion. Crocker & Algina (1986, p. 311).

Traditional analyses of the data collected from the HSTEC field test will be viewed using these three perspectives, specifically item difficulty levels and item discrimination indices.

Item difficulty levels are defined as the proportion of students who correctly answered an item. Identified as p it shows whether or not the item was easy or difficult for the field test students; the higher the p -value the easier the item and conversely, the lower the p -value the more difficult the item.

Item discrimination indices essentially separate examinees who are relatively high on the pass criterion from those who are relatively low. By using these indices one can identify items that high scoring examinees have a higher probability of answering correctly and lower scoring examinees have a lower probability of answering correctly.

Iteman, the program that was used for the traditional psychometric analyses of the field test data divides the discrimination indices into a discrimination index (Crocker and Algina, 1986) and the point biserial correlational index. Discussion and interpretation of these two indices constitute the traditional psychometric perspective of test construction.

For discrimination indices, Crocker and Algina (1986) prescribe the following:

- If $D > .40$, the item functions satisfactorily
- If $.30 < D < .39$, little or no revision of the item is required.
- If $.20 < D < .29$, the item is marginal and needs revision.
- If $D < .19$, the item should be eliminated or completely revised.

Consistent with supporting opinions on the variability of point biserial correlation values (Gulliksen 1945, Richardson 1936), Thorndike (1971, p 142) states that "the point biserial correlation can be said to be a combined measure of item criterion relationship and of difficulty level. The point biserial correlation therefore is not invariant with change in difficulty level." In essence, since the point biserial correlation is related to item difficulty, the point biserial correlation value changes as ability level of the group changes.

Unlike the prescriptiveness of the D statistic in the item difficulty index, interpretation of point biserial correlation lacks such clarity. For free response items Thorndike (1971, p. 152) is specific in that an average biserial correlation value of .6 or .7 defines the test to be homogeneous (with p -value ranges .84 to .16) and will discriminate well across all levels of achievement. A test with

average item test biserial correlations of .3 or .4 would be heterogeneous (with p-values ranges .60 to .40) and not discriminate as well as the homogeneous example previously. For dicotomous items, in order to account for guessing, however, Lord (1953) gives the following summary: "optimum measurement of a given examinee's ability by means of multiple-choice items requires an item difficulty level somewhat easier than halfway between the chance success level and 1.00". For the HSTEC field test, with this in mind and with four alternatives, the average proportion correct should be approximately .75. For interpretation of the point biserial correlation, there appears to be no guidelines for appropriate values. For local test development purposes, the critical value for the point biserial correlation was considered to be .40.

With these guidelines in mind and using results of the Iteman program, the following portion of the paper will discuss data resulting from the Iteman report.

Initial analyses were conducted on an Iteman report formatted for 60 items over 6 subtests/competencies. The results are shown in Table 1.

Insert Table 1 here.

Discussion Using CTT

Following Crocker's recommendation for revision, four items are targeted as having $D < .29$ (3, 6, 23, and 26). Percent correct for these four items are all $> .90$ showing that the discrimination powers for these easy items do not appear to contribute to the purpose of the test (i.e. discriminate for competency), although easy items are needed in the final revision. Using .40 as a guideline for the critical value of the point biserial correlation items 2, 27, 40, and 44 were targeted for further study. The p-values for these four items are $< .55$; D values ranged from .31 to .50. Further investigation into item content and subtest focus showed that item idiosyncrasies or the examinee's background variables rather than instructional experiences could have contributed to their unusual statistics.

Using the traditional psychometric guidelines (item difficulty, p-value, and point biserial correlation), items 2, 3, 6, 23, 26, 27, 40, and 44 would be targeted as needing serious revision or possible elimination from subsequent field testing. However, one needs to keep in mind that, "the nature of the construct being measured may be altered if items are selected purely on the basis of statistical interest without regard to the initial test specifications (Crocker, p 336)."

Item Response Theory (IRT) Methodology

Item response theory (IRT) methodology evolved as a response to concerns with major shortcomings in classical test theory (CTT). In classical test theory examinee characteristics and test characteristics could not be separated, i.e., each had to be interpreted in the context of the other (Hambleton, 1991). In

particular, item indices' values depended on the ability of the group of examinees and examinees' ability levels depended on the choice of items that were selected for the test. In addition, according to Lord (1984), the assumption of equal error of measurement is not plausible and created problems with the interpretation of reliability. Lastly, CTT did not provide predictive powers associated with the generalizability of the test nor did it provide significance tests for model fit.

With these CTT weaknesses in mind, item response theory (IRT) is founded on basic postulates:

- performance of an examinee on a test item can be predicted or explained by a set of factors called traits, latent traits, or abilities and
- the relationship between examinees' item performance and the set of traits underlying item performance can be described by an item characteristic curve (ICC), a function that specifies that as the level of the trait increases, the probability of a correct answer to an item increases.

The assumption of IRT unidimensionality (local independence) for HSTEC field test items holds for the HSTEC field test administration in that the very basis of a competency test assumes that the examinee does have a latent trait, i.e. ability associated with item performance. Further, after conditioning for examinees' ability levels, no relationship exists between examinees' responses to different items (Hambleton, 1991, p.9).

The RASCAL program, using unconditional maximum likelihood (based on the number right) for estimating ability parameters and item difficulty levels, was used to analyze the field test data from an IRT one parameter model perspective. Assuming that only the item difficulty varies (all items are assumed equally discriminating [a_i]) b_i (item difficulty parameter) will be the point on the θ scale where the probability of a correct response will be 0.50. The greater the value of b_i the greater the ability that is required for an examinee to have a .50 chance of getting the item right. Further, values of b_i near -2.0 show items to be easy and those near +2.0 relatively difficult (Camilli and Shepard, 1994). In the one parameter model, like the two and three parameter models, the lower asymptote is zero and is the point at which examinees with low ability levels have zero probability of answering the item correctly. Hambleton (1991, p.30) emphasizes that these assumptions are appropriate for criterion referenced tests following effective instruction.

Discussion Using IRT

RASCAL options for reporting includes standardizing both for ability and for item difficulty. For our purposes, the discussion will focus on standardizing item difficulty since the criterion of interest is examinee ability (i.e. competence).

After only 4 iterations the RASCAL reported the data as shown in Table 2 (df = 19).

Insert Table 2 here.

By comparing data from Table 3 and Table 1, ordinal parameters that describe item difficulty are consistent. The X^2 test for model fit, (Table 4), however, gives information relative to both examinee ability and item difficulty levels. With $df=19$, $X^2 = 30.144$, 37% of the items are described as inappropriate for the Rasch model fit. These discrepancies appear to have no pattern (i.e., falling into one category, e.g. easy), however, there were noticeably more questionable items in the Fact/Opinion (6), Reading (5), and Math (4) subtests. If the HSTEC's intents and purposes are kept in mind, the items that do not fit the Rasch model can be studied as they relate to item difficulty, test construction and competency. Test validity and reliability also enter into the decision of item revision/deletion. For example, items 2, 40, and 44, in combination with other considerations, were:

- earmarked in CTT as needing revision/deletion
- the more difficult items
- identified with very high X^2 values.

After discussion and considering Crocker's caution in not using only statistical significance to gauge item worth, items 2 and 40 would be deleted, but item 44 would remain intact in future versions of the HSTEC.

Insert Table 3 here.

Insert Table 4 here.

Insert Table 5 here.

Table 6 shows that 55% of the students who were administered this field test were able to successfully pass 73% of the items. Converted graphically, (Item by Person Distribution Map) the field test proved to have items normally distributed and examinees who were of higher ability levels. Using the predictive qualities inherent in θ values, and knowing the bias in this particular sample, tests can be designed locally for a more normally distributed population.

Insert Table 6 here.

Conclusion

From this study we can conclude that using both Classical Test Theory (CTT) and Item Response Theory (IRT) during the test development process will more likely result in an instrument that provides data that are reliable to make inferences that are valid. With increased emphases being placed on student achievement scores in accountability systems, it is critical during the test development process to incorporate as much data as feasible in designing an instrument that is based on multiple sources of information. The value that each perspective contributes towards test development further supports the notion that no one method or philosophy should be used in developing tools to measure latent traits. More importantly and in consonance with this analysis, consideration for whether or not the item tests what is valued, should receive primary emphasis.

Figure 1
Hawaii State Test of Essential Competencies

BASIC Skills	Essential Competency	Minimum Score	Maximum Score	BASIC Skills Subtest Passing Score
Reading	EC 1	4	10	48 (passing a minimum of 4 items in each EC and getting 70% of the items correct)
Filling-Out Forms	EC 2	4	10	
Writing	EC 3	4	10	
Oral Language	EC 4	4	9	
Mathematics	EC 5	4	10	
Measurement	EC 6	4	10	
Visual Symbols	EC 6	4	10	
OTHER LIFE Skills				OTHER LIFE Skills Subtest Passing Score
Reasoned Solutions	EC 8	4	10	63 (passing a minimum of 4 items in each EC and getting 70% of the items correct)
Fact vs. Opinion	EC 9	4	10	
Resources	EC 10	4	10	
Health	EC 11	4	10	
Work and Careers	EC 12	4	10	
Government	EC 13	4	10	
Voting	EC 14	4	10	
Citizenship	EC 15	4	10	
Global Diversity	EC 16	4	10	

Table 1 Item Statistics (ITEMAN)

Item No.	Scale-Item	p-value	D Statistic	R _{pbi}
1	1-1	.47	.49	.46
2	1-2	.31	.31	.33
3	1-3	.95	.13	.50
4	1-4	.79	.37	.46
5	1-5	.61	.60	.51
6	1-6	.93	.19	.56
7	1-7	.88	.31	.58
8	1-8	.49	.60	.52
9	1-9	.78	.34	.44
10	1-10	.82	.41	.56
11	2-1	.68	.64	.59
12	2-2	.53	.53	.45
13	2-3	.56	.80	.64
14	2-4	.81	.50	.58
15	2-5	.73	.58	.57
16	2-6	.73	.58	.56
17	2-7	.72	.66	.63
18	2-8	.56	.69	.57
19	2-9	.78	.52	.56
20	2-10	.75	.55	.54
21	3-1	.63	.66	.55
22	3-2	.61	.67	.55
23	3-3	.92	.27	.68
24	3-4	.84	.43	.61
25	3-5	.86	.36	.59
26	3-6	.90	.28	.62
27	3-7	.55	.46	.40
28	3-8	.84	.38	.59
29	3-9	.89	.32	.62
30	3-10	.90	.30	.65
31	4-1	.76	.42	.48
32	4-2	.42	.88	.43
33	4-3	.79	.47	.60
34	4-4	.86	.69	.65
35	4-5	.62	.64	.57
36	4-6	.77	.49	.61
37	4-7	.76	.52	.63
38	4-8	.83	.40	.64
39	4-9	.60	.63	.55
40	4-10	.40	.50	.38
41	5-1	.63	.70	.61
42	5-2	.80	.42	.59
43	5-3	.78	.53	.60
44	5-4	.35	.38	.34
45	5-5	.87	.41	.68
46	5-6	.68	.61	.57
47	5-7	.79	.49	.63
48	5-8	.56	.73	.57

49	5-9	.86	.38	.66
50	5-10	.74	.53	.57
51	6-1	.59	.55	.52
52	6-2	.67	.60	.58
53	6-3	.49	.61	.54
54	6-4	.53	.61	.54
55	6-5	.42	.52	.49
56	6-6	.41	.52	.47
57	6-7	.74	.56	.60
58	6-8	.41	.47	.47
59	6-9	.59	.63	.57
60	6-10	.75	.50	.57

Table 2 Final Parameter Estimates (RASCAL)

Item No.	Difficulty [b _i]	SE	X ²
1	1.28	.08	52.364
2	2.11	.08	166.064
3	-2.67	.20	23.321
4	-0.52	.10	51.545
5	0.59	.08	30.037
6	-2.20	.17	18.364
7	-1.38	.12	32.710
8	1.18	.08	20.794
9	-0.93	.09	53.670
10	-0.76	.10	16.578
11	0.23	.08	45.093
12	0.99	.08	67.255
13	0.83	.08	15.106
14	-0.64	.10	17.973
15	-0.08	.09	23.234
16	-0.06	.09	44.143
17	-0.00	.09	32.190
18	0.86	.08	25.727
19	-0.40	.09	12.768
20	-0.19	.09	13.993
21	0.49	.08	19.942
22	0.62	.08	13.697
23	-1.95	.15	31.241
24	-0.93	.11	29.892
25	-1.14	.11	28.468
26	-1.66	.14	17.262
27	0.91	.08	79.908
28	-0.87	.11	20.499
29	-1.46	.13	17.601
30	-1.66	.14	22.844
31	-0.26	.09	31.272
32	1.56	.08	50.343
33	-0.48	.10	46.470
34	-1.08	.11	42.697

35	0.53	.08	28.757
36	-0.32	.09	42.275
37	-0.25	.09	25.274
38	-0.77	.10	31.198
39	0.65	.08	22.288
40	1.64	.08	63.053
41	0.48	.08	28.691
42	-0.73	.10	25.573
43	-0.44	.09	24.725
44	1.91	.08	115.809
45	-1.20	.12	74.356
46	0.21	.08	22.899
47	-0.51	.10	23.359
48	0.83	.08	25.331
49	-1.14	.11	27.668
50	-0.12	.09	17.661
51	0.71	.08	34.556
52	0.25	.08	13.793
53	1.49	.08	25.764
54	1.02	.08	31.766
55	1.55	.08	19.420
56	1.61	.08	22.795
57	-0.12	.09	25.833
58	1.61	.08	29.677
59	0.72	.08	14.703
60	-0.18	.09	19.533

Table 3 Final Parameter Estimates (RASCAL) by difficulty levels (difficult to easy)

Item No.	Difficulty [b _i]	SE	X ²
2	2.11	.08	166.064
44	1.91	.08	115.809
40	1.64	.08	63.053
56	1.61	.08	22.795
58	1.61	.08	29.677
32	1.56	.08	50.343
55	1.55	.08	19.420
53	1.49	.08	25.764
1	1.28	.08	52.364
8	1.18	.08	20.794
54	1.02	.08	31.766
12	0.99	.08	67.255
27	0.91	.08	79.908
18	0.86	.08	25.727
13	0.83	.08	15.106
48	0.83	.08	25.331
59	0.72	.08	14.703
51	0.71	.08	34.556
39	0.65	.08	22.288

22	0.62	.08	13.697
5	0.59	.08	30.037
35	0.53	.08	28.757
21	0.49	.08	19.942
41	0.48	.08	28.691
52	0.25	.08	13.793
11	0.23	.08	45.093
46	0.21	.08	22.899
17	-0.00	.09	32.190
16	-0.06	.09	44.143
15	-0.08	.09	23.234
50	-0.12	.09	17.661
57	-0.12	.09	25.833
60	-0.18	.09	19.533
20	-0.19	.09	13.993
37	-0.25	.09	25.274
31	-0.26	.09	31.272
36	-0.32	.09	42.275
19	-0.40	.09	12.768
43	-0.44	.09	24.725
33	-0.48	.10	46.470
47	-0.51	.10	23.359
4	-0.52	.10	51.545
14	-0.64	.10	17.973
42	-0.73	.10	25.573
10	-0.76	.10	16.578
38	-0.77	.10	31.198
28	-0.87	.11	20.499
9	-0.93	.09	53.670
24	-0.93	.11	29.892
34	-1.08	.11	42.697
25	-1.14	.11	28.468
49	-1.14	.11	27.668
45	-1.20	.12	74.356
7	-1.38	.12	32.710
29	-1.46	.13	17.601
26	-1.66	.14	17.262
30	-1.66	.14	22.844
23	-1.95	.15	31.241
6	-2.20	.17	18.364
3	-2.67	.20	23.321

Table 4 Final Parameter Estimates (RASCAL) by X^2

Item No.	Difficulty [b]	SE	X^2
2	2.11	.08	166.064
44	1.91	.08	115.809
27	0.91	.08	79.908
45	-1.20	.12	74.356
12	0.99	.08	67.255
40	1.64	.08	63.053
9	-0.93	.09	53.670
1	1.28	.08	52.364
4	-0.52	.10	51.545
32	1.56	.08	50.343
33	-0.48	.10	46.470
11	0.23	.08	45.093
16	-0.06	.09	44.143
34	-1.08	.11	42.697
36	-0.32	.09	42.275
51	0.71	.08	34.556
7	-1.38	.12	32.710
17	-0.00	.09	32.190
54	1.02	.08	31.766
31	-0.26	.09	31.272
23	-1.95	.15	31.241
38	-0.77	.10	31.198
5	0.59	.08	30.037
24	-0.93	.11	29.892
58	1.61	.08	29.677
35	0.53	.08	28.757
41	0.48	.08	28.691
25	-1.14	.11	28.468
49	-1.14	.11	27.668
57	-0.12	.09	25.833
53	1.49	.08	25.764
18	0.86	.08	25.727
42	-0.73	.10	25.573
48	0.83	.08	25.331
37	-0.25	.09	25.274
43	-0.44	.09	24.725
47	-0.51	.10	23.359
3	-2.67	.20	23.321
15	-0.08	.09	23.234
46	0.21	.08	22.899
30	-1.66	.14	22.844
56	1.61	.08	22.795
39	0.65	.08	22.288
8	1.18	.08	20.794
28	-0.87	.11	20.499
21	0.49	.08	19.942
60	-.18	.09	19.533
55	1.55	.08	19.420

6	-2.20	.17	18.364
14	-0.64	.10	17.973
50	-0.12	.09	17.661
29	-1.46	.13	17.601
26	-1.66	.14	17.262
10	-0.76	.10	16.578
13	0.83	.08	15.106
59	0.72	.08	14.703
20	-0.19	.09	13.993
52	0.25	.08	13.793
22	0.62	.08	13.697
19	-0.40	.09	12.768

Table 5 Item Statistics (ITEMAN) sorted by p-value

Item No.	Scale-Item	p-value	D Statistic	R _{phi}
3	1-3	.95	.13	.50
6	1-6	.93	.19	.56
23	3-3	.92	.27	.68
26	3-6	.90	.28	.62
30	3-10	.90	.30	.65
29	3-9	.89	.32	.62
7	1-7	.88	.31	.58
45	5-5	.87	.41	.68
25	3-5	.86	.36	.59
34	4-4	.86	.69	.65
49	5-9	.86	.38	.66
24	3-4	.84	.43	.61
28	3-8	.84	.38	.59
10	1-10	.82	.41	.56
14	2-4	.81	.50	.58
38	4-8	.8	.40	.64
42	5-2	.80	.42	.59
4	1-4	.79	.37	.46
33	4-3	.79	.47	.60
47	5-7	.79	.49	.63
9	1-9	.78	.34	.44
19	2-9	.78	.52	.56
43	5-3	.78	.53	.60
36	4-6	.77	.49	.61
31	4-1	.76	.42	.48
37	4-7	.76	.52	.63
20	2-10	.75	.55	.54
60	6-10	.75	.50	.57
50	5-10	.74	.53	.57
57	6-7	.74	.56	.60
15	2-5	.73	.58	.57
16	2-6	.73	.58	.56
17	2-7	.72	.66	.63
11	2-1	.68	.64	.59
46	5-6	.68	.61	.57

52	6-2	.67	.60	.58
21	3-1	.63	.66	.55
41	5-1	.63	.70	.61
35	4-5	.62	.64	.57
5	1-5	.61	.60	.51
22	3-2	.61	.67	.55
39	4-9	.60	.63	.55
51	6-1	.59	.55	.52
59	6-9	.59	.63	.57
13	2-3	.56	.80	.64
18	2-8	.56	.69	.57
48	5-8	.56	.73	.57
27	3-7	.55	.46	.40
12	2-2	.53	.53	.45
54	6-4	.53	.61	.54
8	1-8	.49	.60	.52
53	6-3	.49	.61	.54
1	1-1	.47	.49	.46
32	4-2	.42	.88	.43
55	6-5	.42	.52	.49
56	6-6	.41	.52	.47
58	6-8	.41	.47	.47
40	4-10	.40	.50	.38
44	5-4	.35	.38	.34
2	1-2	.31	.31	.33

Table 6 Raw Score Conversion (RASCAL)

Number Correct	Ability (Theta)	SE	Frequency	Percentile	Scaled Score
0	+++++	+++++	16	2	+++
1	-4.55	1.03	0	2	58
2	-3.91	0.74	0	2	64
3	-3.46	0.62	0	2	68
4	-3.13	0.54	0	2	72
5	-2.87	0.49	0	2	74
6	-2.64	0.46	0	2	76
7	-2.45	0.43	0	2	78
8	-2.27	0.41	0	2	79
9	-2.11	0.39	1	2	81
10	-1.97	0.38	0	2	82
11	-1.83	0.36	1	2	83
12	-1.70	0.35	1	2	85
13	-1.58	0.34	0	2	86
14	-1.47	0.34	2	3	87
15	-1.35	0.33	3	3	88
16	-1.25	0.32	1	3	89
17	-1.15	0.32	5	4	90
18	-1.05	0.31	5	4	90

19	-0.95	0.31	4	5	91
20	-0.86	0.30	2	5	92
21	-0.77	0.30	5	6	93
22	-0.64	0.30	4	6	94
23	-0.59	0.30	11	7	95
24	-0.50	0.29	5	8	95
25	-0.41	0.29	6	9	96
26	-0.33	0.29	7	10	97
27	-0.24	0.29	10	11	98
28	-0.16	0.29	12	12	99
29	-0.07	0.29	12	14	99
30	0.01	0.29	17	16	100
31	0.09	0.29	15	18	101
32	0.18	0.29	18	20	102
33	0.26	0.29	14	22	102
34	0.35	0.29	11	23	103
35	0.43	0.29	17	25	104
36	0.52	0.29	18	27	105
37	0.61	0.30	27	31	106
38	0.69	0.30	17	33	106
39	0.78	0.30	23	35	107
40	0.87	0.30	34	40	108
41	0.97	0.31	29	43	109
42	1.06	0.31	36	47	110
43	1.16	0.32	28	51	111
44	1.26	0.32	31	55	111
45	1.37	0.33	36	59	112
46	1.47	0.33	44	64	113
47	1.59	0.34	29	68	114
48	1.70	0.35	34	72	116
49	1.83	0.36	34	76	117
50	1.96	0.37	35	81	118
51	2.11	0.39	33	85	119
52	2.26	0.40	32	89	121
53	2.43	0.43	42	94	122
54	2.62	0.45	20	96	124
55	2.84	0.49	20	99	126
56	3.10	0.54	18	99	128
57	3.42	0.61	3	99	131
58	3.86	0.74	5	99	135
59	4.59	1.03	2	99	142
60	++++	+++++	1	99	+++

References

- Camili, G. & Shepard, L.A. (1994). Methods for Identifying Biased Test Items. Thousand Oaks, CA: Sage.
- Crocker, L.; & Algina, J. (1986). Introduction to Classical and Modern Test Theory. Orlando, FL: Holt, Rinehart, and Winston.
- Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). Fundamentals of item response theory. Newbury Park, CA: Sage.
- Nunnaly, J. & Koplin, J.H (1967). The effects of word-relatedness on learning. ERIC Document Reproduction Service ED 016 214.
- Lord, F.M. (1984). Maximum Likelihood and Bayesian Parameter Estimation in Item Response Theory. ERIC Document Reproduction Service ED 250 365.
- Thorndike, R. L. (1971). Concepts of culture-fairness. Journal of Educational Measurement, 8, 63-70.